SUPPLEMENTAL MATERIAL

Human Subjects and DNA Samples

Written informed consent for genetic studies was obtained in agreement with protocols approved by the institutional review boards (IRB) at Vanderbilt University and Columbia University Medical Center. DNA was extracted from peripheral blood leukocytes using Puregene reagents (Gentra Systems Inc., Minnesota, USA).

Exome Capture

The Agilent SureSelect Human All Exon Kit adapted for the SOLiD sequencing platform was applied. Genomic DNA (5 μg) was randomly fragmented by sonication, treated with Klenow fragment of DNA polymerase I (NEB) to generate blunt ends and then phosphorylated with polynucleotide kinase (NEB). Adaptor primers were annealed and ligated to the fragment ends. Ligated samples were hybridized with the baits for 48 hours, washed and eluted using the Agilent protocol. Following elution, the capture efficiency was evaluated via q-PCR reactions.

Next Generation Sequencing and Analysis

The resulting captured DNA samples were subjected to standard sample preparation procedures for SOLiD sequencing according to the manufacturer's instructions.

The analysis pipeline included genome alignment followed by SNP/indel variant analysis. BWA (Burrows-Wheeler Aligner) [1] was used to align 50-base reads to human genome GRCh37, allowing for a 3% error rate, 2 gaps and maximum edit distance of 5 bases. A custom module was used to select reads that most likely associate with the captured regions. SAMtools [2] was

used to call targeted bases, with valid-adjacent base calls that deviate from the reference treated as potential variations (SNP/indel) and assigned a coverage-dependent Phred-scaled mutation probability. Called SNPs and indels had minimum depth of 10 reads, and variation quality, a Phred-scaled probability that the consensus is identical to the reference, of up to 50% for indels and 20% for SNPs.

We sequenced four samples (II-4, III-2, III-5, IV-2; Figure 1) with an average of 87.3M reads per sample. Of the 334M sequenced reads, 97.6% had more that 45% present color calls per 50-base read (mappable reads) and were aligned to the human genome; 35% of these reads could be aligned uniquely to the genome and 62% of the uniquely aligned reads aligned to a SureSelect region, resulting in a 26x mean coverage of 88% of the SureSelect regions. Theses 71M reads included 34% duplicate reads, which is within two standard deviations from expectation according to simulations of starting-point selection from SureSelect regions at this level of coverage, suggesting that the majority of our replicates are not PCR induced. We used the filtering framework by Sasson et al. [18] to estimate sequencing errors in mappable reads, but removed reads only based on alignment quality. In total, 20% of the 326M reads had an average read quality value above 20; 74% had a high quality leading base call (quality value above 25 in one of the first 10 positions in the read), and 47% had at least 3 high quality leading base calls. Only 18% of reads had a high quality leading base call and fewer than 6 color call errors. After alignment, we used reads that mapped uniquely to the SureSelect regions to predict variations. We identified 54,540 genomic locations with variations in any of the 4 samples, and 10,088 (18%) of these locations had variations shared in all 4 samples. We identified 27,908 variations in patient II-4, 28,279 variations in patient III-2, 23,846 variations in patient III-5, and 22,483

variations in patient IV-2 with 2679 – 4041 novel SNPs in each subject with the expected ratio of transversions to transitions (Supplemental Table 1).

The single nucleotide and indel variants were compared among the four individual DNAs. There were 653 heterozygous variants shared among all four family members that were screened using dbSNP (hg18), SIFT, and the 1000 Genomes Project. Novel variants present in all four patients were analyzed for effect on the amino acid sequence. Nonsynonymous variants, splice mutations, and coding insertions/deletions for each subject were tabulated and were further evaluated (Supplemental Table 2). There were 52 novel variants present in all four patients, with 31 in coding regions. Of the 31 novel coding variants, 16 were predicted to cause nonsynonymous changes. These 16 variants were sequenced using dideoxy sequencing and 11 were confirmed. The 11 confirmed variants were in the following genes: *FKBP4, ZHX3, OR1Q1, PASK, ADAMTSL3, OGT2, PRSS1, F2RL1, TNFAIP2, FEM1B, and CAV1*. The five variants that were not confirmed were found to have similar sequence in the genome that led to mis-alignment using the shotgun approach with short reads. We then used DNA from an additional affected family member IV-1 (who was not exon sequenced) and attempted to confirm the 11 non-synonymous coding variants. Only 3 of these non-synonymous coding variants were also carried by IV-1. These three variants were S36T in olfactory receptor family 1 subfamily Q member 1 *(OR1Q1)*, H379Y coagulation factor II receptor like 1 (*F2RL1*), and P158PfsX22 in Caveolin-1 (*CAV1)*. When we analyzed these variants using SNAP, [19] the S36T *OR1Q1* and the H379Y *F2RL1* variants were predicted to have neutral effects with a reliability index of 2 and 4 and an expected accuracy of 69 and 85%, respectively. The c.474delA in *CAV1* was predicted to cause

a frameshift P158P fsX22 and add 21 novel amino acids at the C terminal domain of CAV1 protein.

Capillary Sequencing

For follow-up confirmation of identified novel variants, dideoxy sequencing was applied. PCR primers were designed flanking approximately 250 bp of a given variant and sequenced on ABI 3730 capillary sequencing instrument. Capillary sequence reads were analyzed using the Sequencher software package (GeneCodes Inc).

Lung Tissue Sampling

Lung biopsy was performed for clinical purposes on one patient diagnosed with idiopathic pulmonary hypertension that had an identified novel mutation in *CAV1*. Lung tissue was obtained at open-lung biopsy and was fixed in 10% formalin, processed, embedded in paraffin, sectioned and stained with hematoxylin-eosin, CD31, alpha smooth muscle actin (SMA) and Verhoeff-Van Gieson *(*VVG*)* Elastic Staining.

REFERENCES

1. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 2009;25:1754-60.

2. Li H, Handsaker B, Wysoker A, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics 2009;25:2078-9.

3. Zuber TJ. Punch biopsy of the skin. Am Fam Physician 2002;65:1155-8, 61-2, 64.

Supplemental Table 1. Summary of the variants identified in each of the subjects with WES including number of homozygous and heterozygous variants, the ratio of heterozygotes to homozygotes, the number of variant transitions and transversions, and the ratio of transversions to transitions (Tv/Ts). The expected ratio of transversions to transitions within the region captured is 0.43 in 1000 genomes and in agreement with the ratio observed in our subjects.

ID	Total SNP	Homo	Hetero	Het/Hom ratio	Transitions	Transversion	Tv/Ts ratio	Total SNPs	Known SNPs	Novel SNPs	% novel	Novel/ Known
PPH48HM2141	27,907	13,504	14,403	1.0666	18,923	7,232	0.38	27907	24300	3607	12.9	0.15
PPH48KL1144	23,845	11,785	12,060	1.0233	16,202	6,171	0.33	23845	20801	3044	12.8	0.15
PPH48SW1367	28,278	13,648	14,630	1.0720	18,892	7,601	0.40	28278	24237	4041	14.3	0.17
PPH48RW99	22,482	10,869	11,613	1.0685	15,079	5,932	0.31	22482	19803	2679	11.9	0.14

Supplemental Table 2. Functional annotation of number and frequency of variants of different types for the four subjects with WES.

Functional Annotation	PPH48HM2141		PPH48KL1144		PPH48SW1367		PPH48RW99	
CODON_CHANGE	0	0.00%	0	0.00%	35	0.07%	23	0.06%
CODON_CHANGE_PLUS_CODON_DELETION	18	0.04%	11	0.03%	20	0.04%	14	0.03%
CODON_CHANGE_PLUS_CODON_INSERTION	13	0.03%	12	0.03%	16	0.03%	13	0.03%
CODON_DELETION	7	0.01%	6	0.01%	5	0.01%	4	0.01%
CODON_INSERTION	12	0.02%	11	0.03%	12	0.02%	17	0.04%
DOWNSTREAM	3,085	5.93%	2,463	5.51%	3,000	5.67%	2,103	5.03%
FRAME_SHIFT	275	0.53%	279	0.62%	257	0.49%	226	0.54%
INTERGENIC	3,167	6.08%	2,657	5.95%	3,104	5.87%	2,387	5.71%
INTRAGENIC	907	1.74%	813	1.82%	831	1.57%	691	1.65%
INTRON	25,452	48.90%	21,713	48.60%	26,317	49.73%	21,211	50.73%
NONE	112	0.22%	53	0.12%	68	0.13%	53	0.13%
NON_SYNONYMOUS_CODING	6,738	12.95%	6,096	13.64%	7,046	13.31%	5,435	13.00%
NON_SYNONYMOUS_START	2	0.00%	2	0.00%	2	0.00%	1	0.00%
SPLICE_SITE_ACCEPTOR	20	0.04%	16	0.04%	37	0.07%	29	0.07%
SPLICE_SITE_DONOR	50	0.10%	42	0.09%	44	0.08%	41	0.10%
START_GAINED	47	0.09%	41	0.09%	38	0.07%	32	0.08%
START_LOST	9	0.02%	15	0.03%	10	0.02%	7	0.02%
STOP_GAINED	49	0.09%	50	0.11%	60	0.11%	43	0.10%
STOP_LOST	5	0.01%	7	0.02%	7	0.01%	9	0.02%
SYNONYMOUS_CODING	8,363	16.07%	7,328	16.40%	8,444	15.96%	6,505	15.56%
SYNONYMOUS_STOP	5	0.01%	6	0.01%	6	0.01%	4	0.01%
UPSTREAM	1,952	3.75%	1,563	3.50%	1,825	3.45%	1,532	3.66%
UTR_3_PRIME	1,285	2.47%	1,077	2.41%	1,257	2.38%	1,052	2.52%
UTR_5_PRIME	471	0.91%	416	0.93%	479	0.91%	375	0.90%

CODON_CHANGE: One or many codons are changed

CODON_CHANGE_PLUS_CODON_DELETION : A deletion multiple of three at codon boundary

CODON_CHANGE_PLUS_CODON_INSERTION : An insert multiple of three in a codon boundary

CODON_DELETION: Deletion of a codon

CODON_INSERTION: Insertion of a codon

DOWNSTREAM: Downstream of a gene in 5000 bps

FRAME_SHIFT: Resulting in a frameshift

INTERGENIC: Variant in a intergenic region

INTRAGENIC: The variant is within a gene, but no transcripts within the gene

INTRON: Variant in an intron

NONE: All functional annotations that don't fall into any of the functional categories (including all variants larger than a point mutation)

NON_SYNONYMOUS_CODING: Non-synonymous coding variant

NON_SYNONYMOUS_START: Variant causes a codon that produces a different amino acid and variant causes start codon to be mutated into another start codon

SPLICE_SITE_ACCEPTOR: Variant in a splice acceptor site

SPLICE_SITE_DONOR: Variant in a splice donor site

START_GAINED: Variant creating a novel start site

START_LOST: Variant in the start site

STOP_GAINED: Variant creating a novel termination

STOP_LOST: Variant in the termination sequence

SYNONYMOUS_CODING: Synonymous variant in the coding sequence

SYNONYMOUS_STOP: Synonymous variant in the termination site

UPSTREAM: Upstream of a gene in 5000 bps

UTR_3_PRIME: Variant in the 3’ untranslated region

UTR_5_PRIME: Variant in the 5’ untranslated region